# Accident Severity Prediction

**IBM Capstone Project**

Bibhash Chandra Mitra

# Introduction

There are many factors which play a role in any type of accident. We cannot pinpoint a single reason behind the severity of an accident. But out of all those factors there are some which can be tracked and are definite in nature. For example, the weather condition, light condition, type of sidewalk etc are some of the factors which are deterministic. Other factors like number of pedestrians, speeding etc. are not deterministic without a high speed camera so we won't mention them. So can we say that using these factors we can answer questions like: What is the probability of an accident occurring on a rainy day at an Intersection?

I believe we can answer these type of questions and hence the purpose of this project is to devise a solution for the prediction of severe car accidents based on the different conditions which can be effectively tracked.

# Data Preparation

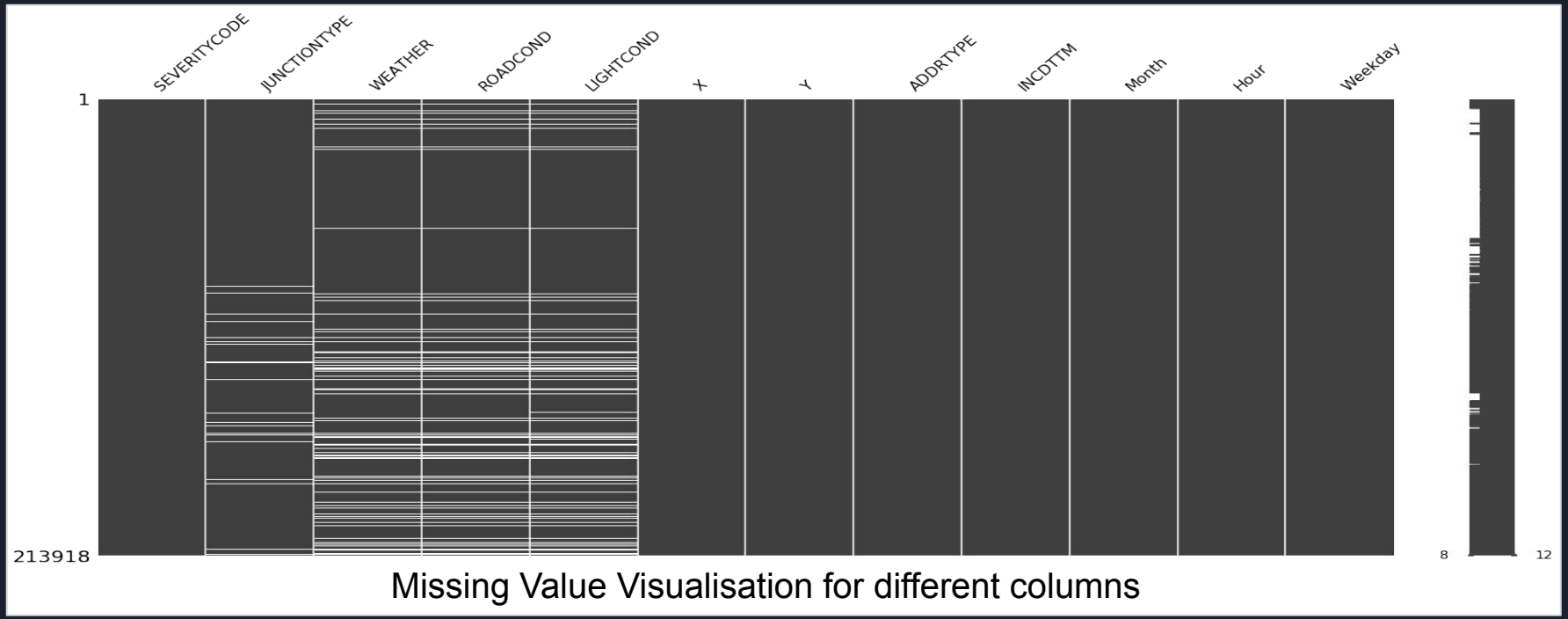
For this project I have used the Seattle transit Collision dataset. This dataset contains all the type of collisions from 2004 to present. There are 38 columns in this dataset, out of those we will use the following:

- **JUNCTIONTYPE** : Category of junction at which collision took place
- **WEATHER** : A description of the weather conditions during the time of the collision.
- **ROADCOND** : The condition of the road during the collision
- **LIGHTCOND** : The light conditions during the collision.
- **INCDTTM** : The date and time (From which we can get information on the month and time of the day)
- **X** : Longitude of the location
- **Y** : Latitude of the location
- **ADDRTYPE** : Collision address type
  - Alley
  - Block
  - Intersection
- **SEVERITYCODE** : A code that corresponds to the severity of the collision:
  - 3—fatality
  - 2b—serious injury
  - 2—injury
  - 1—prop damage
  - 0—unknown

# Data Preprocessing

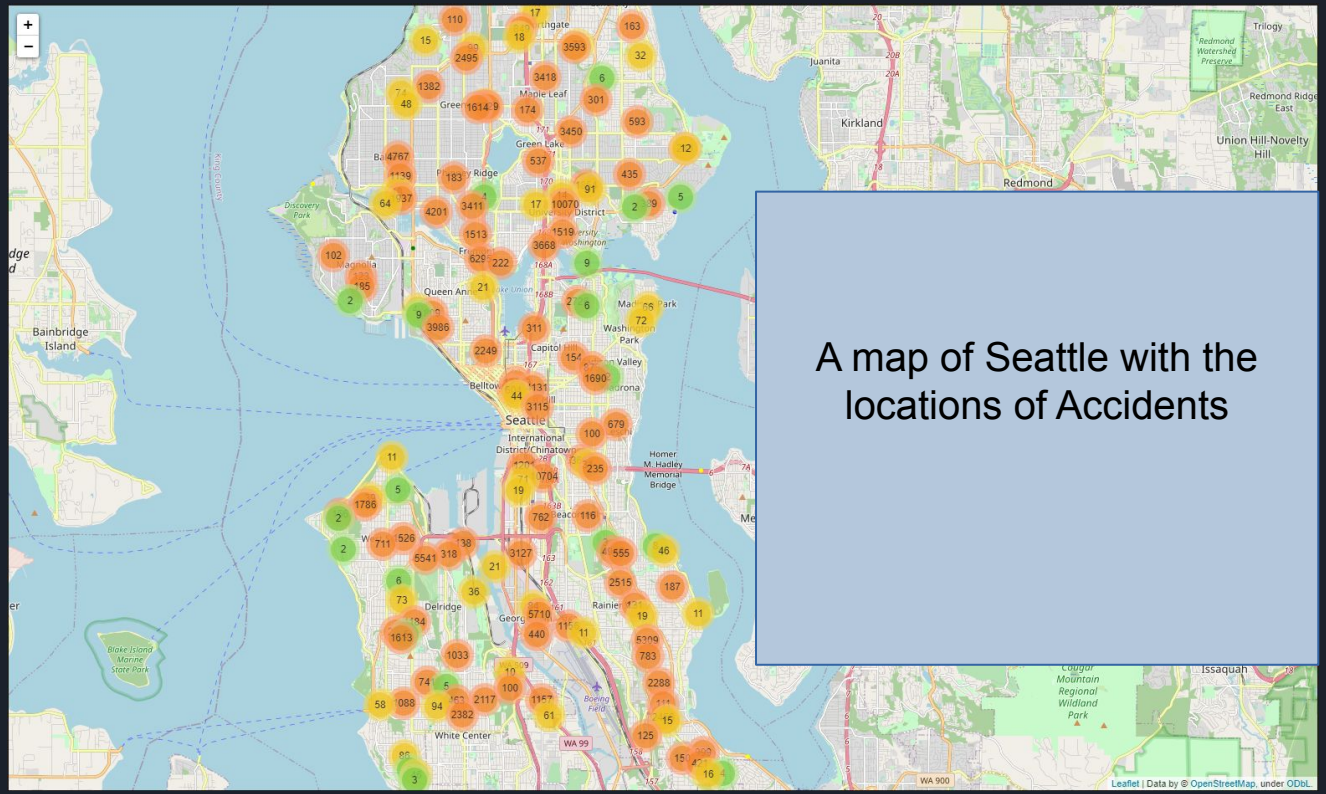
- Feature Selection
- Modifying and Adding Features
- Handling the missing values
- One Hot Encoding.
- Sampling the Data and balancing

# Missing values



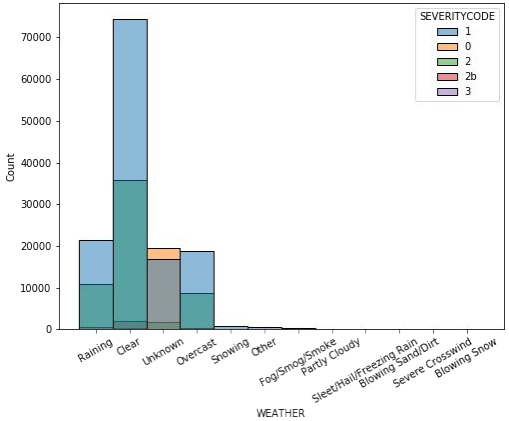

Missing Value Visualisation for different columns

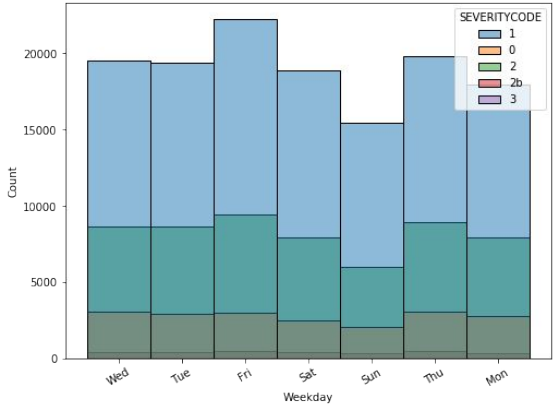
# Data Visualisation

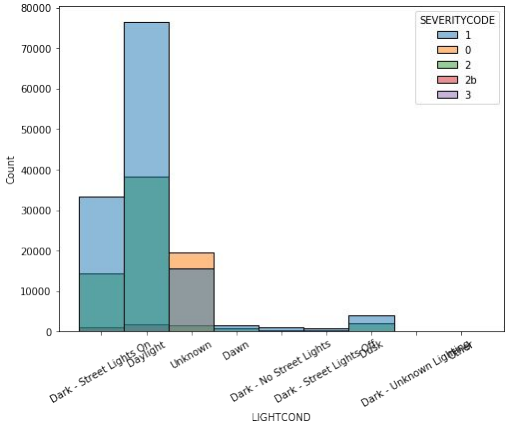


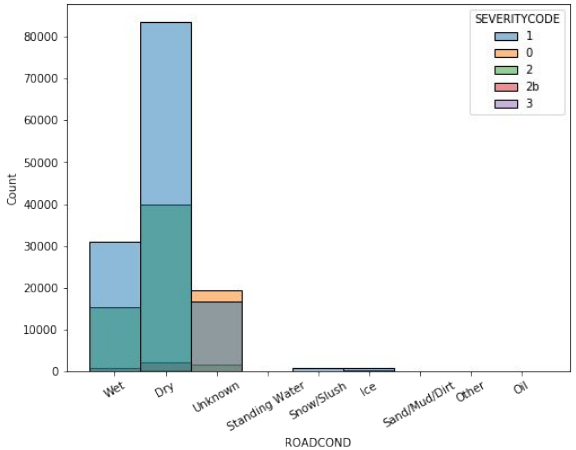

A map of Seattle with the locations of Accidents

# Data Visualisation



# Results and Discussion

Based on the various results from different algorithms used as listed below, we were able to achieve an F1-score of 62 percent which is not very reliable in real life scenarios. But we can make further improvements using more feature engineering.

The algorithms are as follows:
1. KNN Algorithm with k=5
2. Decision Trees with max_depth=6
3. Random Forest with n_estimators=300

| Algorithm | Jaccard | F1-score |
|---|---|---|
| KNN | 0.57 | 0.57 |
| Decision Tree | 0.62 | 0.62 |
| Random Forest | 0.58 | 0.58 |

Our Analysis mainly demonstrates the highest F1-score is achieved using Decision Tree classifier.
This also suggests that we can use some better feature engineering to improve upon our results.

# Conclusion

We can conclude that we can get an accuracy of more than 60 percent using a Decision tree model and can further improve upon the results using better features. So some future scopes are:

- Inclusion of as many features as possible by making the newly created instance from already existing one feature for better prediction.
- Weighted XGBoost, Naïve Bayes and other similar models can be implemented instead of resampling the dataset.
- Detailed study of each feature and their correlation with dependent variable Severity will be done.